

# DS 621: Homework 1 (Group3)

## Homework 1 Description

Zach Alexander, Sam Bellows, Donny Lofland, Joshua REgiste, Neil Shah, Aaron Zalik

### Overview

In this homework assignment, you will explore, analyze and model a data set containing approximately 2200 records. Each record represents a professional baseball team from the years 1871 to 2006 inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season.

Your objective is to build a multiple linear regression model on the training data to predict the number of wins for the team. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	Positive Impact on Wins
TEAM_BATTING_2B	Doubles by batters (2B)	Positive Impact on Wins
TEAM_BATTING_3B	Triples by batters (3B)	Positive Impact on Wins
TEAM_BATTING_HR	Homeruns by batters (4B)	Positive Impact on Wins
TEAM_BATTING_BB	Walks by batters	Positive Impact on Wins
TEAM_BATTING_HBP	Batters hit by pitch (get a free base)	Positive Impact on Wins
TEAM_BATTING_SO	Strikeouts by batters	Negative Impact on Wins
TEAM_BASERUN_SB	Stolen bases	Positive Impact on Wins
TEAM_BASERUN_CS	Caught stealing	Negative Impact on Wins
TEAM_FIELDING_E	Errors	Negative Impact on Wins
TEAM_FIELDING_DP	Double Plays	Positive Impact on Wins
TEAM_PITCHING_BB	Walks allowed	Negative Impact on Wins
TEAM_PITCHING_H	Hits allowed	Negative Impact on Wins
TEAM_PITCHING_HR	Homeruns allowed	Negative Impact on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive Impact on Wins

Figure 1: Variables of Interest

## 1. Data Exploration

Describe the size and the variables in the moneyball training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

- a. Mean / Standard Deviation / Median
- b. Bar Chart or Box Plot of the data
- c. Is the data correlated to the target variable (or to other variables?)
- d. Are any of the variables missing and need to be imputed "fixed"?

### Data Snapshot

**Columns** Descriptive text ... Lorem ipsum dolor sit amet, consectetur adipiscing elit. Fusce ut augue pharetra, luctus lectus ut, rutrum quam. Aenean quis tellus ac felis accumsan pellentesque id ut purus. Fusce eget ligula eu est congue aliquet. Vivamus hendrerit felis varius lorem suscipit venenatis. Fusce facilisis arcu ac lorem cursus, non pretium velit finibus. Suspendisse eu nulla tellus. Nunc viverra elementum dolor, ut scelerisque nisl. Ut iaculis faucibus ultricies. Praesent fermentum eu libero et consequat. Phasellus vitae euismod lectus, a ultrices dui. Nunc vel leo rhoncus, cursus elit quis, rhoncus nisl. Aenean id urna et nibh tempor iaculis nec non quam. In tincidunt luctus ex eget viverra.

### Data Sample

### Dimensions

### Data Summary

### Variable Plots

## 2. Data Preparation

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.

- a. Fix missing values (maybe with a Mean or Median value)
- b. Create flags to suggest if a variable was missing
- c. Transform data by putting it into buckets
- d. Mathematical transforms such as log or square root (or use Box-Cox)
- e. Combine variables (such as ratios or adding or multiplying) to create new variables

### Missing Values

### Outliers

### Impute

### 3. Build Models

Using the training data set, build at least three different multiple linear regression models, using different variables (or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

### 4. Select Models

For the multiple linear regression model, will you use a metric such as Adjusted R<sup>2</sup>, RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R<sup>2</sup>, (c) F-statistic, and (d) residual plots. Make predictions using the evaluation data set.

### References

- A
- B

### Appendix

#### R Code

```
# Show all code here
```